

# 从实验到数据

RNA、文库、fastqc质控和fastp去接头

陈明杰 202410

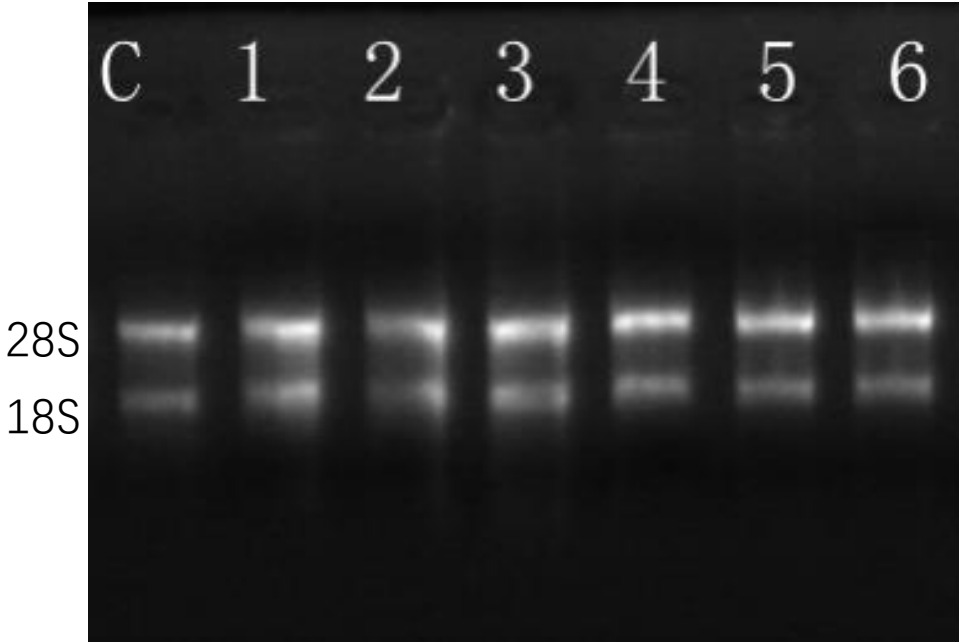
# 样品质控（实验员）

- RNA量
- RNA完整性
- 污染

| <div>上海纽科生物科技有限公司</div> <div>Shanghai NewCore Biotechnology Co., Ltd</div> |           |               |         |            |                               |                                        |                                         |
|----------------------------------------------------------------------------|-----------|---------------|---------|------------|-------------------------------|----------------------------------------|-----------------------------------------|
| RNA样本质检报告                                                                  |           |               |         |            |                               |                                        |                                         |
| 数量                                                                         | 送样单位      | 客户            | 合同号     | 送样日期       | 仓位                            | 样品来源                                   | 样品类型                                    |
| 6                                                                          | ***       | ***           | ***     | ***        | ***                           | ***                                    | ***                                     |
| 电泳顺序                                                                       | 样本名称      | 浓度<br>(ng/μl) | 体积 (μl) | 总量<br>(ng) | 总量评定<br>1.合格<br>2.有风险<br>3.不足 | 大小分布<br>1.合格<br>2.降解<br>3.有杂带<br>4.无杂带 | 综合评定<br>1.合格<br>2.有风险<br>3.不合格<br>4.需纯化 |
| 1                                                                          | cancer-1  | 530.0         | 60      | 31800.0    | 1                             | 1                                      | 1                                       |
| 2                                                                          | cancer-2  | 564.0         | 60      | 33840.0    | 1                             | 1                                      | 1                                       |
| 3                                                                          | cancer-3  | 614.0         | 60      | 36840.0    | 1                             | 1                                      | 1                                       |
| 4                                                                          | control-1 | 554.0         | 60      | 33240.0    | 1                             | 1                                      | 1                                       |
| 5                                                                          | control-2 | 566.0         | 60      | 33960.0    | 1                             | 1                                      | 1                                       |
| 6                                                                          | control-3 | 553.0         | 60      | 33180.0    | 1                             | 1                                      | 1                                       |

浓度检测方法： 荧光染料定量法

标准品



点样孔处有带：蛋白、多糖等污染  
比28S还大的带：基因组DNA污染  
拖尾或弥散：RNA降解

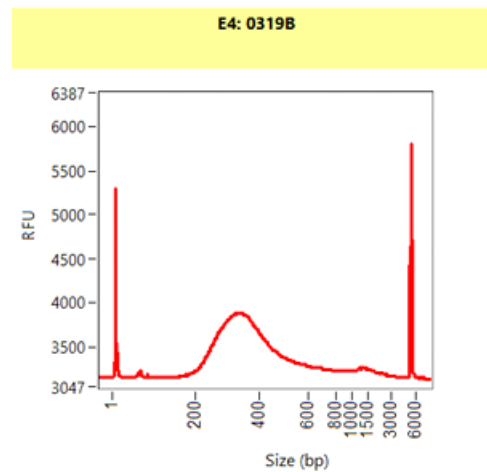
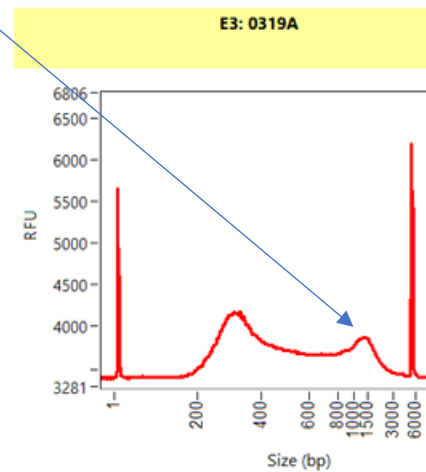
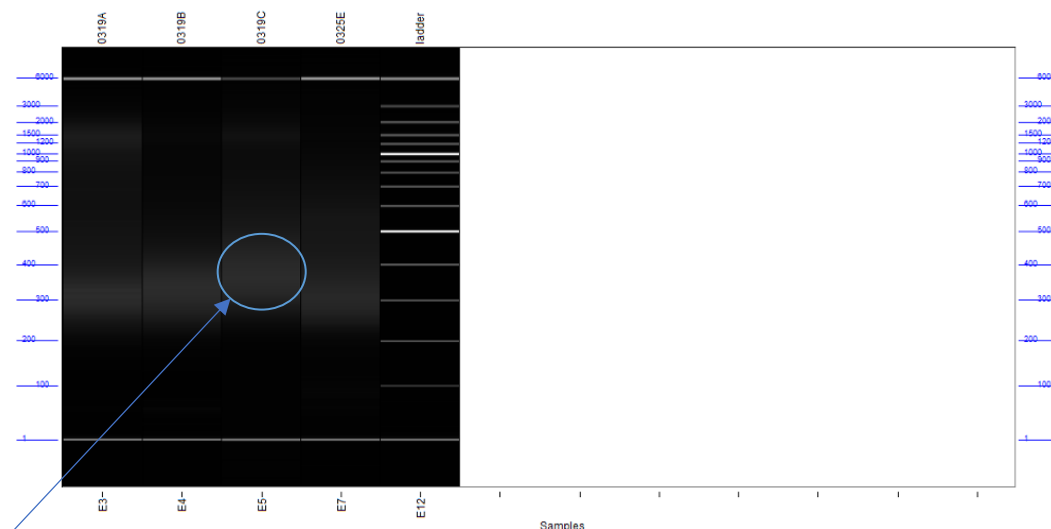
# 文库质控（测序工厂）

- Q-PCR检测文库有效浓度
- Nanodrop 4200检测插入片段大小

1) 文库质量检测结果：

| 序号 | 样本名称  | 样本类型         | Volume (μl) | Qubit (ng/μl) | Qpcr (nM) | 长度 (bp) | 质检结论 | 备注           |
|----|-------|--------------|-------------|---------------|-----------|---------|------|--------------|
| 1  | 0319A | RNA 文库       | 11          | 8.42          | 35.47     | 429.00  | 风险上机 | 总量满足要求，大片段污染 |
| 2  | 0319B | RNA 文库       | 8.5         | 6.04          | 49.62     | 381.00  | 合格   | 总量、片段满足要求    |
| 3  | 0319C | RNA 文库       | 11          | 25            | 115.54    | 437.00  | 风险上机 | 总量满足要求，大片段污染 |
| 4  | 0325D | Small RNA 文库 | 11          | 1.65          | 10.73     | 359.00  | 风险上机 | 总量满足要求，小片段污染 |
| 5  | 0325E | RNA 文库       | 11          | 12.7          | 4.98      | 414.00  | 合格   | 总量、片段满足要求    |

Gel Image

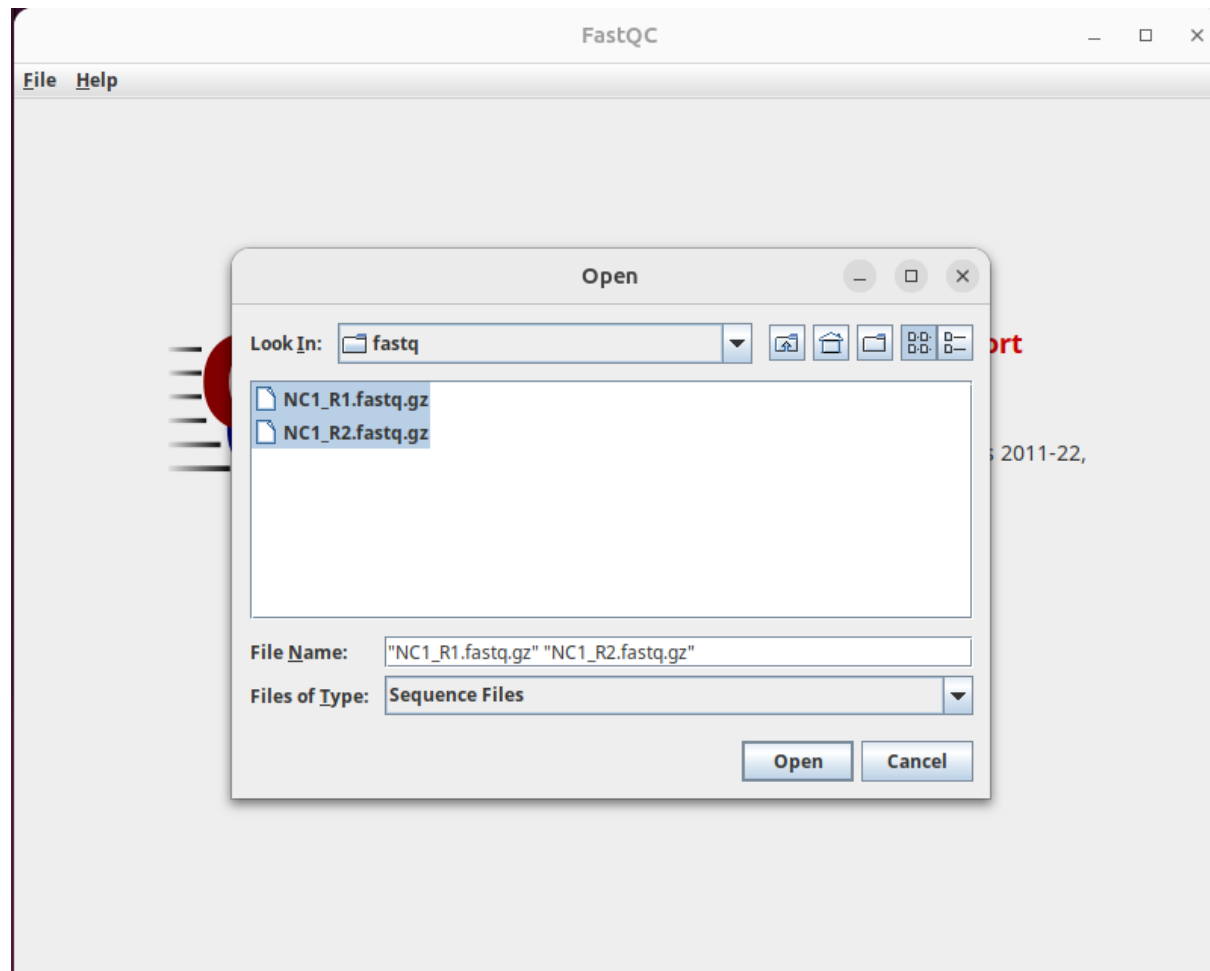
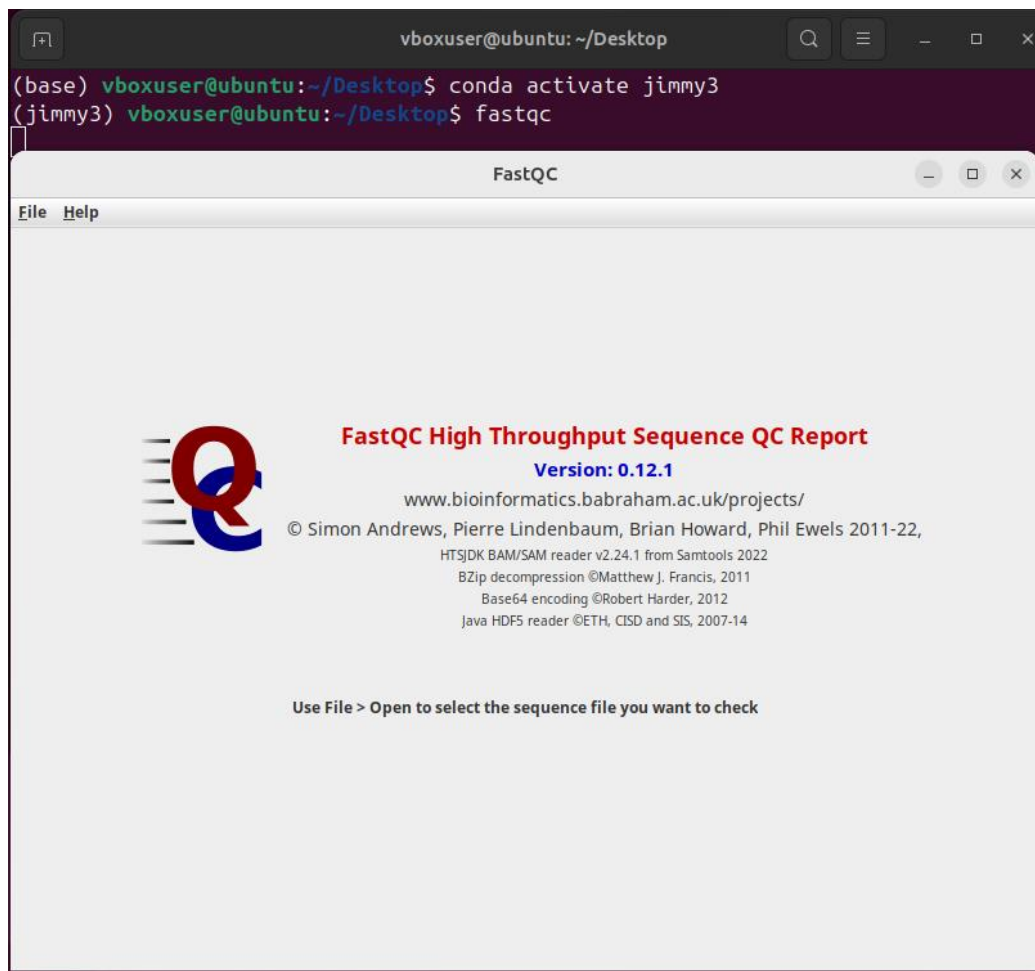


# 下机数据质控（数据分析员）

- 数据量够不够：一般6-8G
- $Q30 > 80\%$ ：技术非常成熟，一般不会出问题（提前拦截）
- 接头比例：一般非常少
- 上机批次（不可控，强调下）：样品少尽量同一条lane上机
- 匹配率：一般90%+。若80以下，加污染检测，例如cell样品可能会有支原体污染

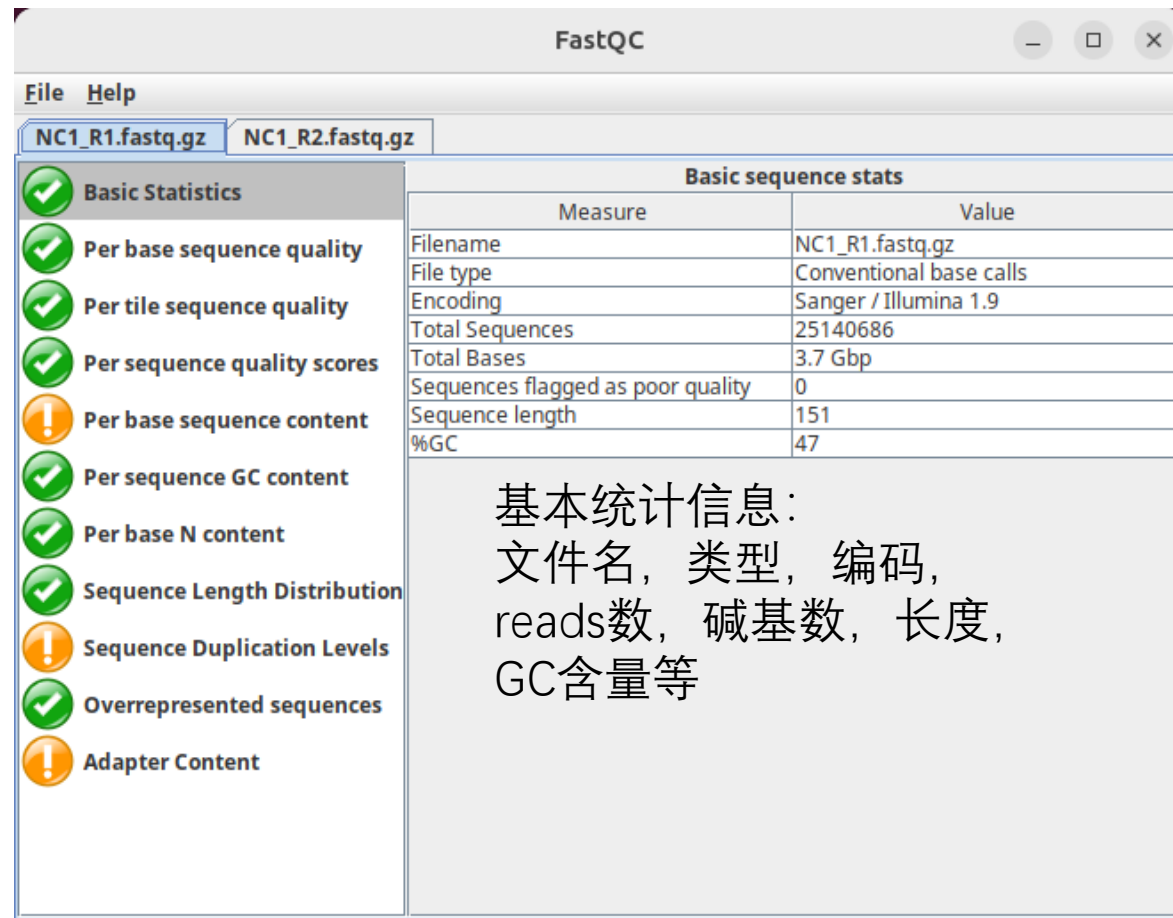
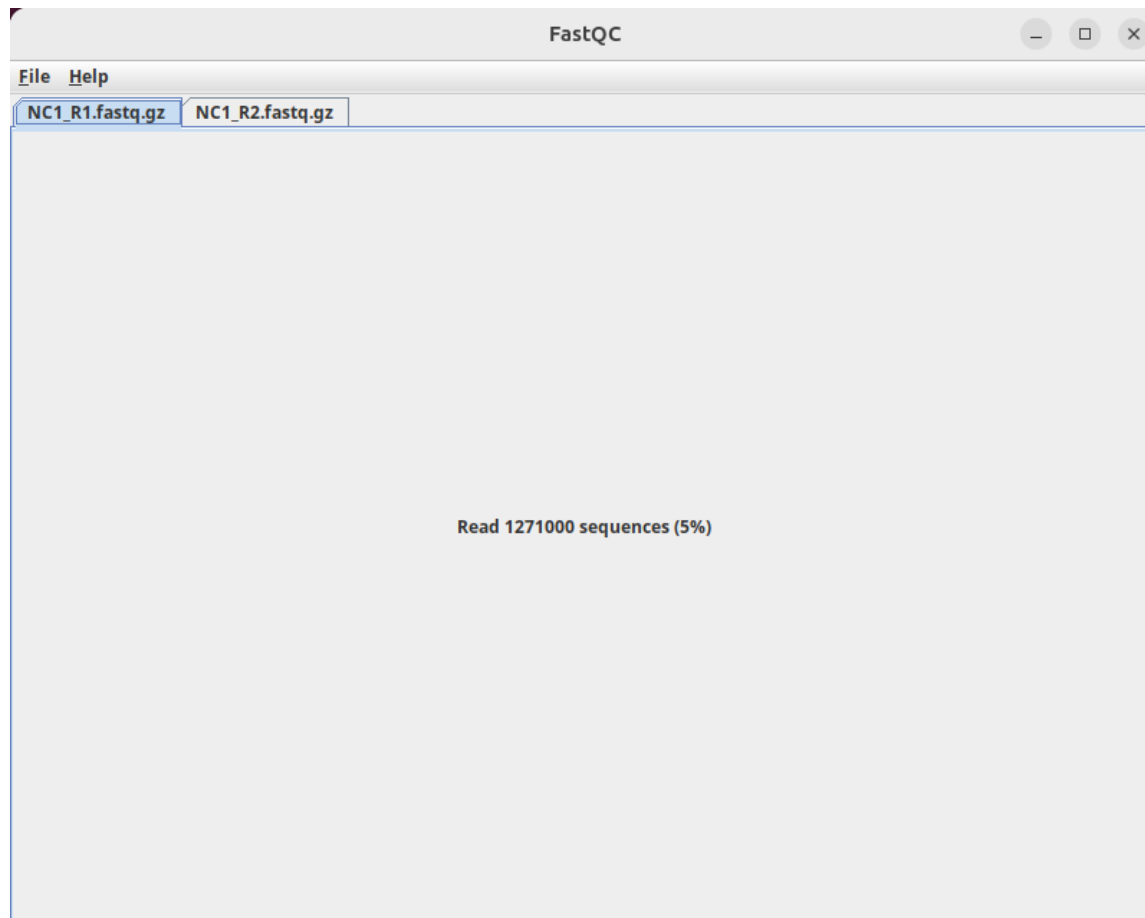
# 打开程序，载入fastq文件

```
fastqc -f fastq -t 10 --noextract *.fastq.gz
```



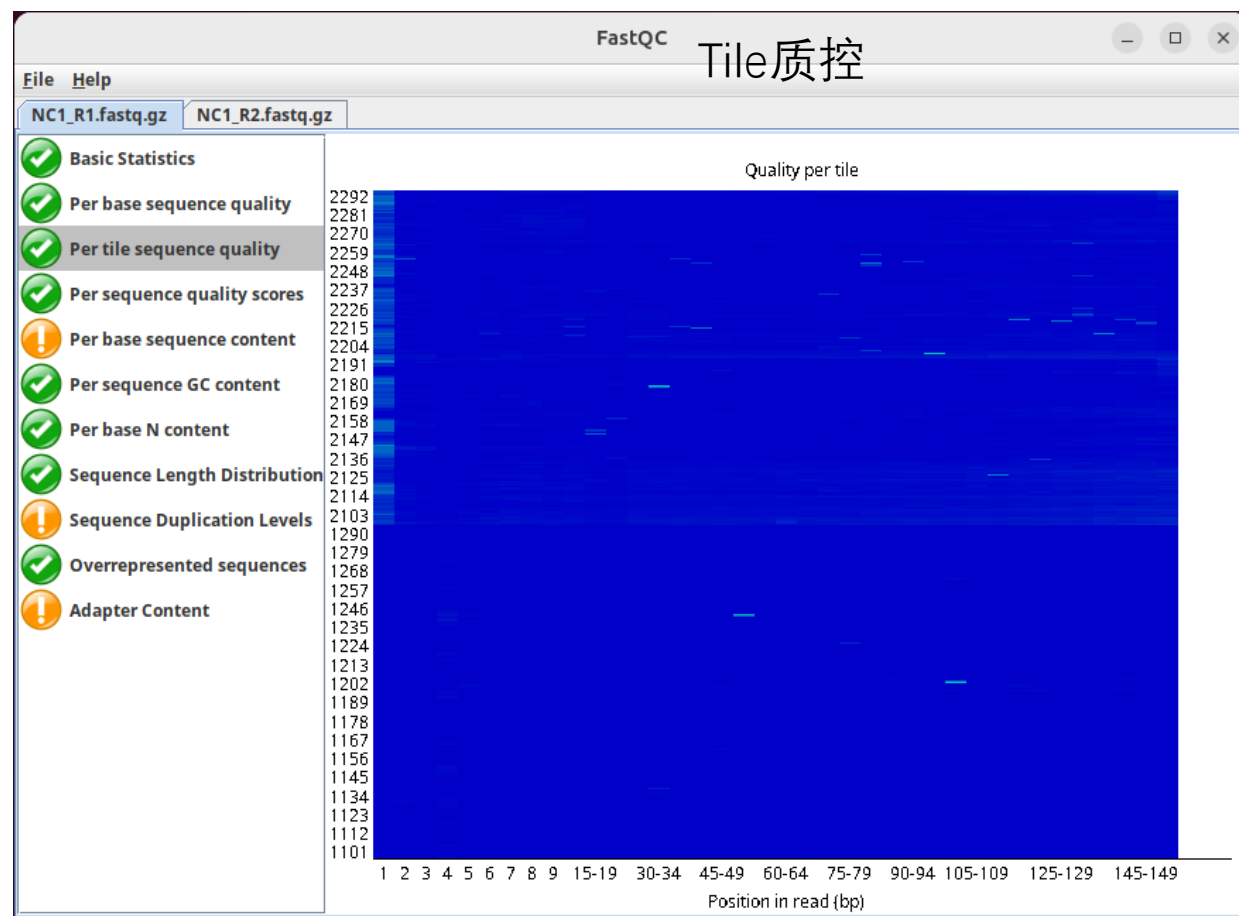
# 依次运行

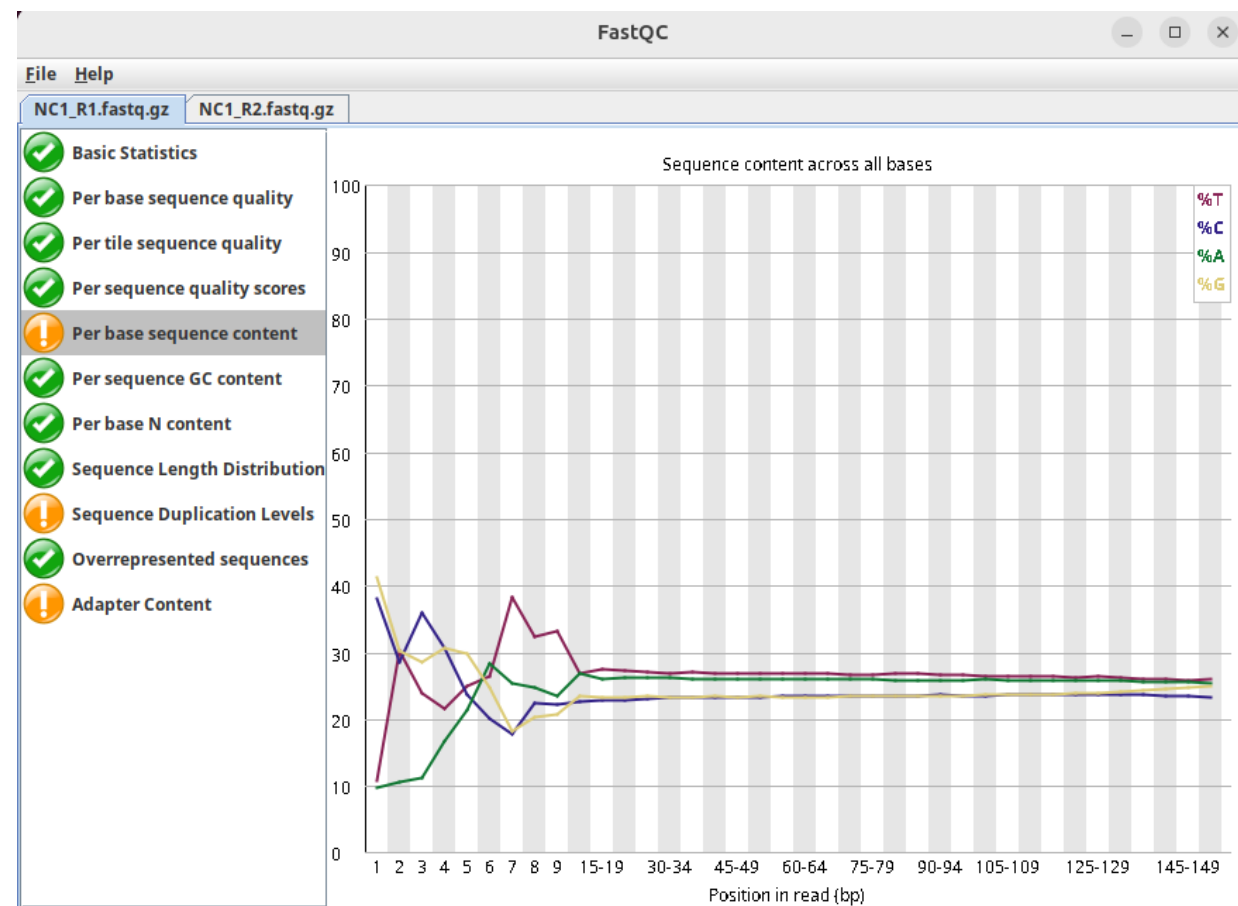
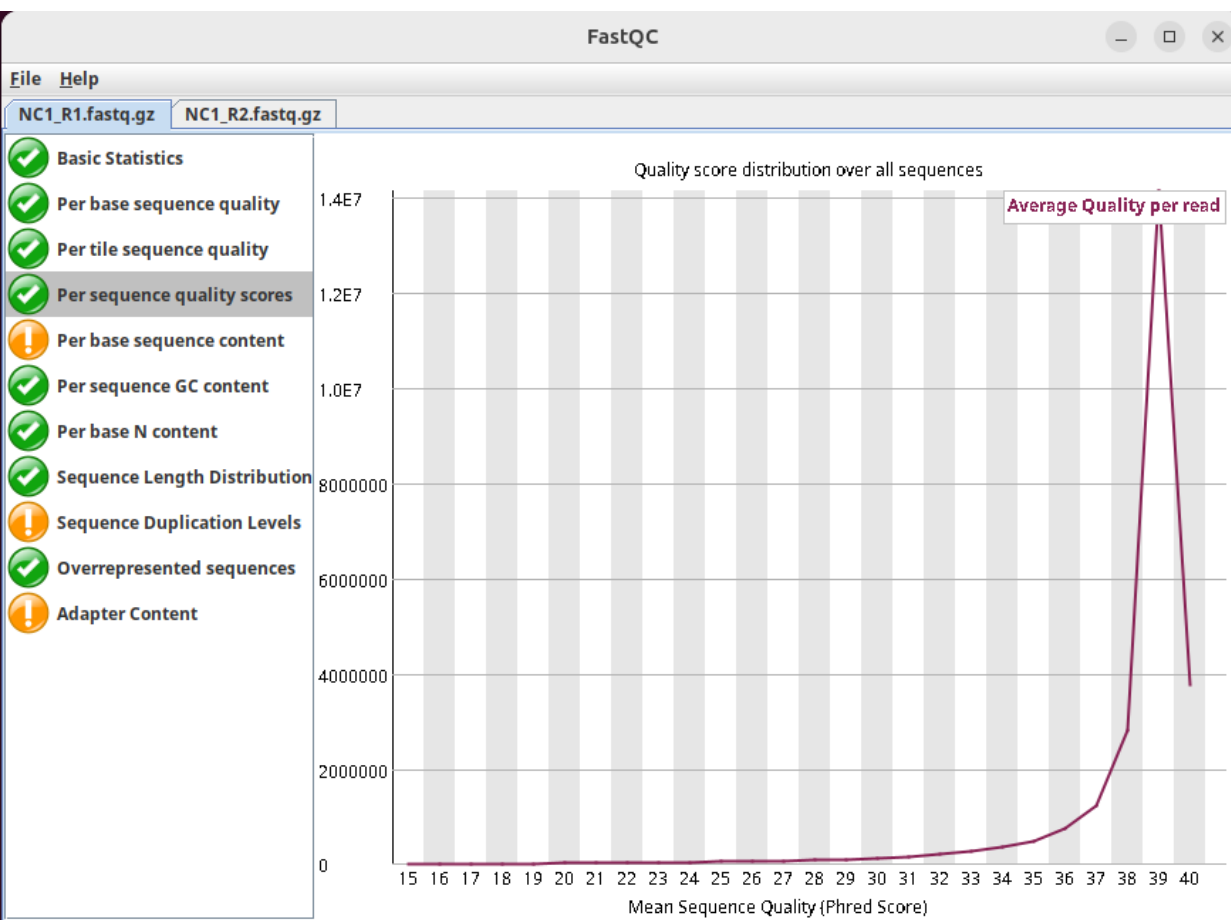
绿色：通过  
 黄色：警告  
 红色：失败



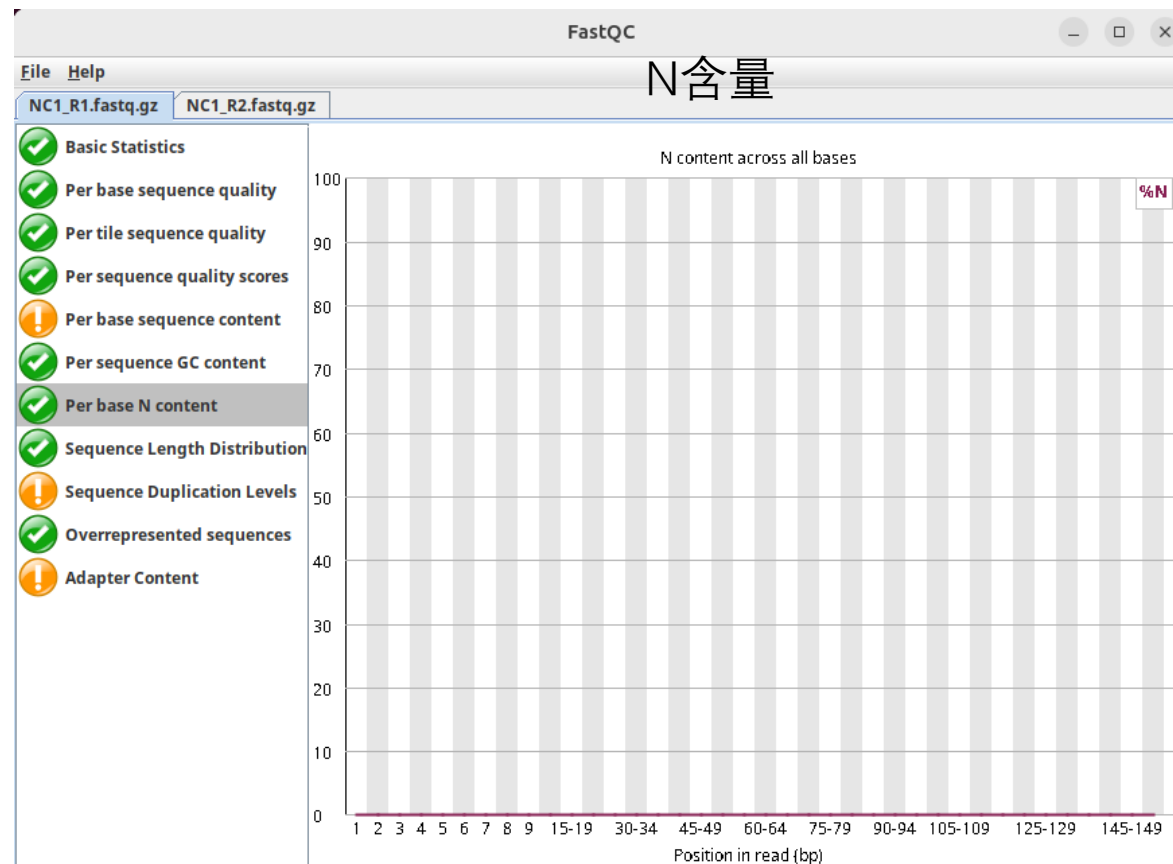
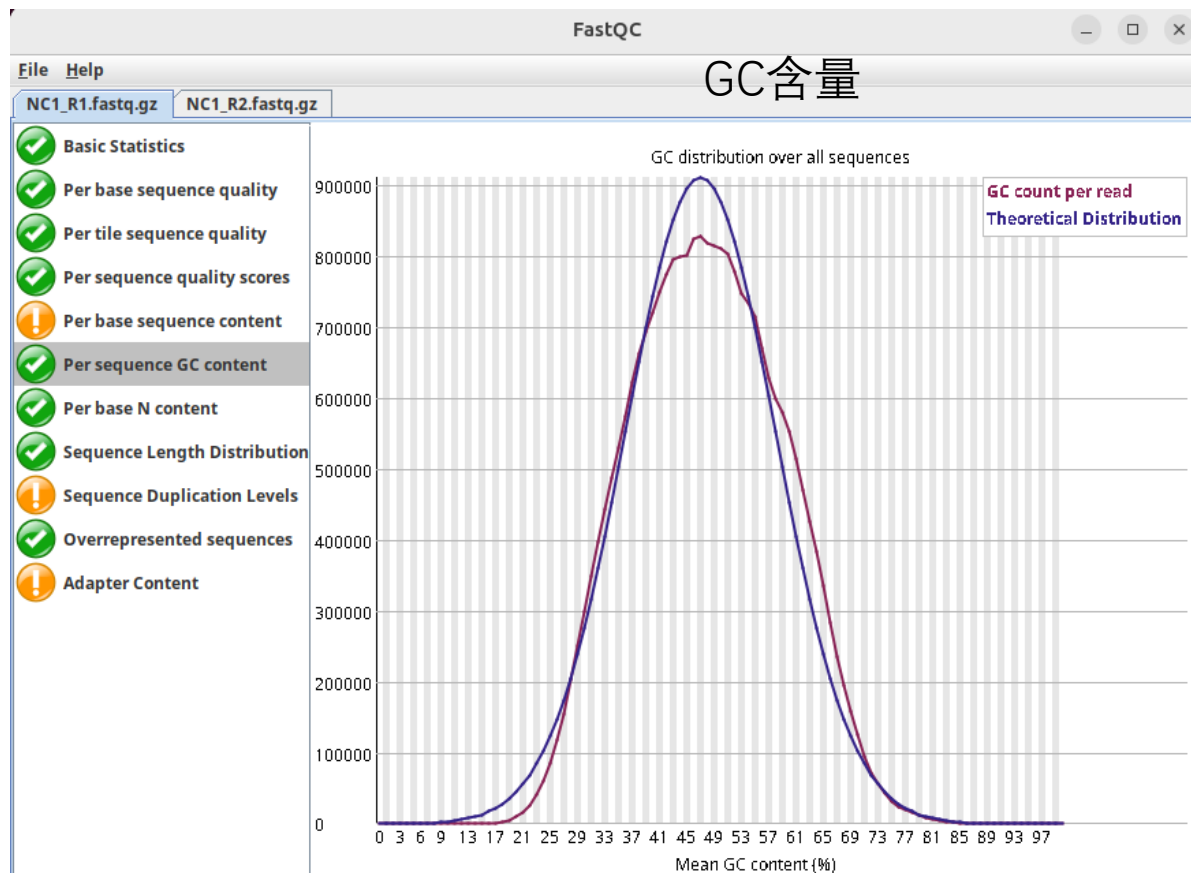


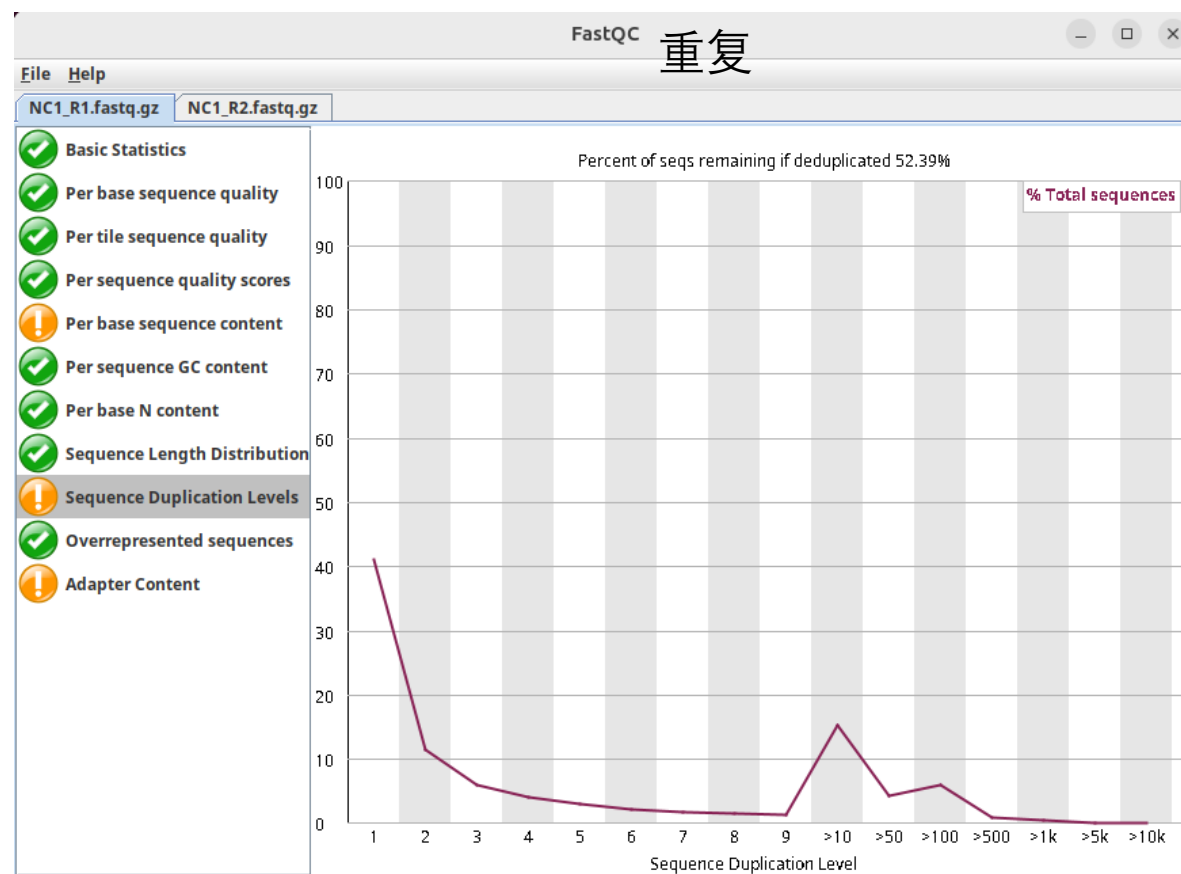
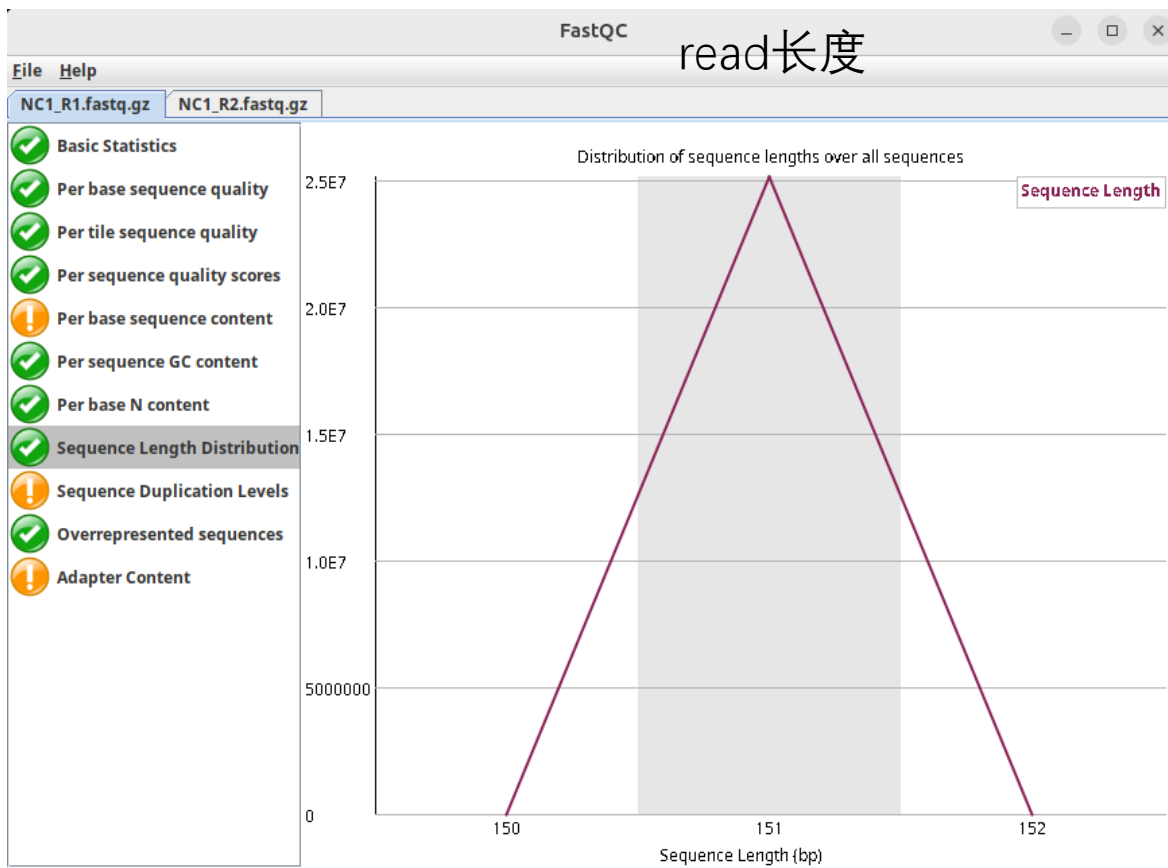
Read位置

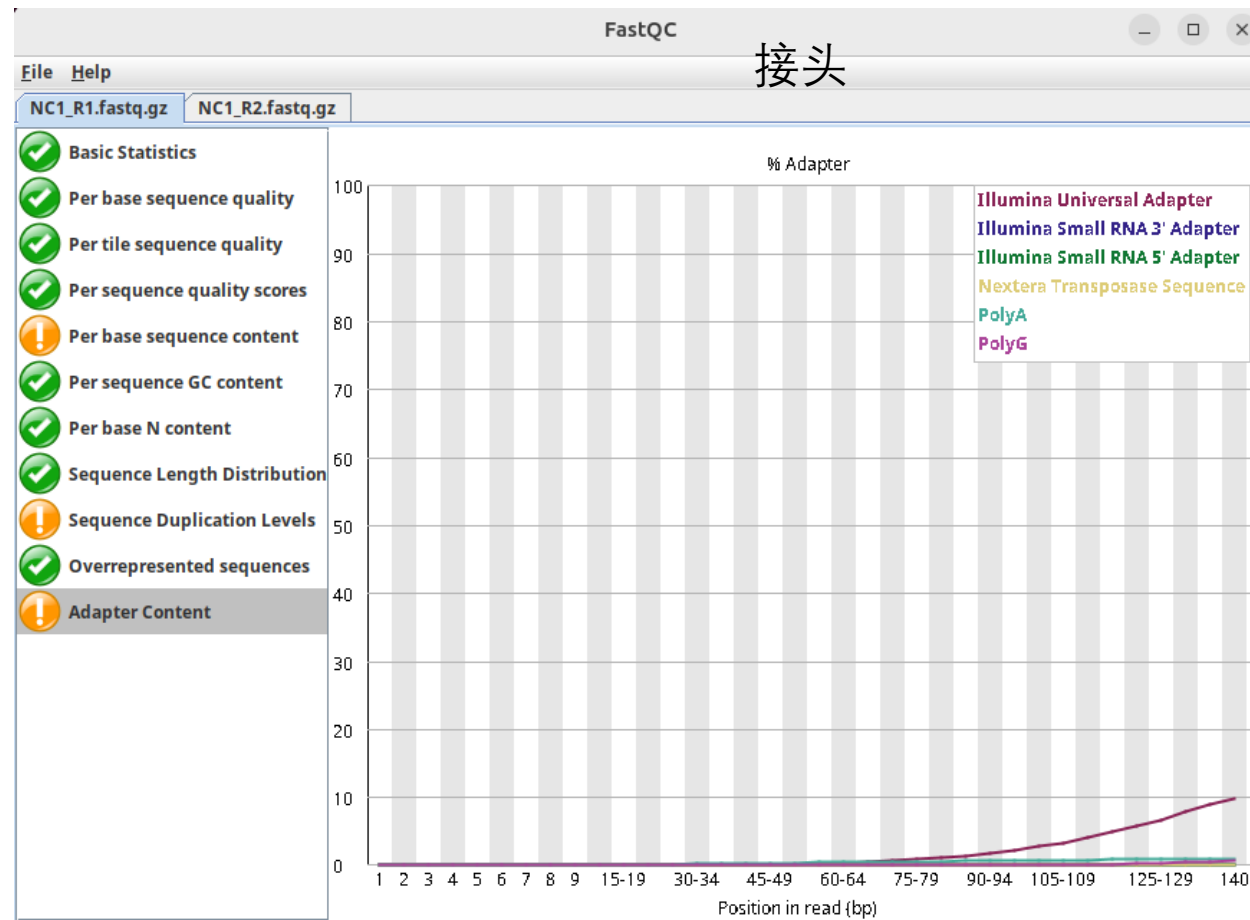












# fastp去接头

```
(jimmy3) vboxuser@ubuntu:~/share/fastq$ fastp -i NC1_R1.fastq.gz -o NC1_R1.clean
.gz -I NC1_R2.fastq.gz -O NC1_R2.clean.gz --detect_adapter_for_pe -w 2 -h NC1.fa
stp.html
Detecting adapter sequence for read1...
>Illumina TruSeq Adapter Read 1
AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
Detecting adapter sequence for read2...
```

```
fastp -i NC1_R1.fastq.gz -o NC1_R1.clean.gz -I NC1_R2.fastq.gz -O NC1_R2.clean.gz --detect_adapter_for_pe -w 2 -h NC1.fastp.html
```

- i read1 fastq输入文件 -o read1输出文件（也是fastq格式）
- I read2 fastq输入文件 -O read2输出文件（也是fastq格式）
- detect\_adapter\_for\_pe 检测双端的接头
- w cpu数
- h 输出的summary html文件名

注：多线程，所以输出文件中reads的顺序跟原来不一样

## Summary

### General

|                               |                                                                                              |
|-------------------------------|----------------------------------------------------------------------------------------------|
| fastp version:                | 0.23.4 ( <a href="https://github.com/OpenGene/fastp">https://github.com/OpenGene/fastp</a> ) |
| sequencing:                   | paired end (151 cycles + 151 cycles)                                                         |
| mean length before filtering: | 151bp, 151bp                                                                                 |
| mean length after filtering:  | 146bp, 146bp                                                                                 |
| duplication rate:             | 6.766908%                                                                                    |
| Insert size peak:             | 194                                                                                          |
| Detected read1 adapter:       | AGATCGGAAGAGCACACGTCTGAACTCCAGTCA                                                            |

### Before filtering

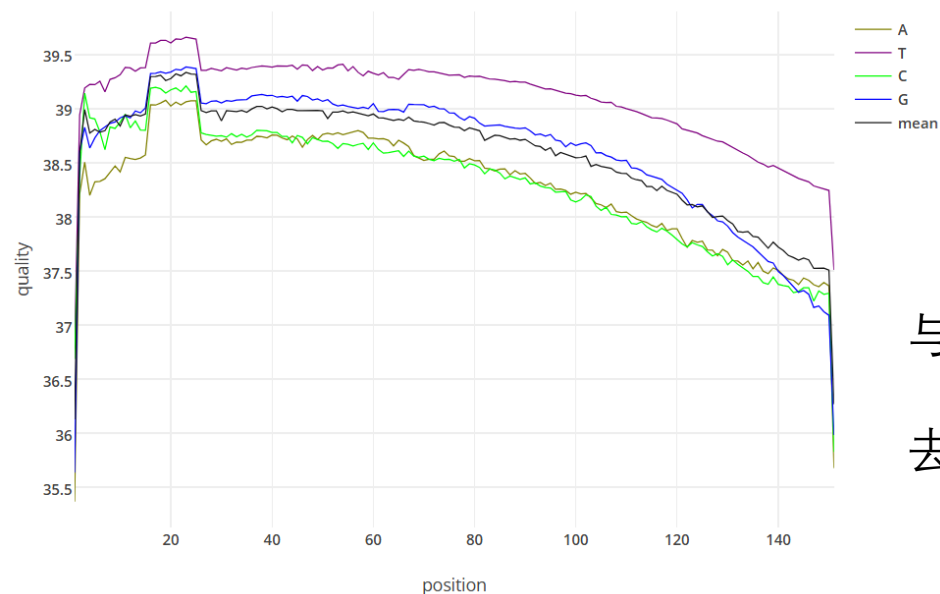
|              |                         |
|--------------|-------------------------|
| total reads: | 50.281372 M             |
| total bases: | 7.592487 G              |
| Q20 bases:   | 7.338749 G (96.658040%) |
| Q30 bases:   | 6.957154 G (91.632086%) |
| GC content:  | 47.979376%              |

### After filtering

|              |                         |
|--------------|-------------------------|
| total reads: | 49.163226 M             |
| total bases: | 7.209235 G              |
| Q20 bases:   | 7.042220 G (97.683324%) |
| Q30 bases:   | 6.734894 G (93.420378%) |
| GC content:  | 47.655670%              |

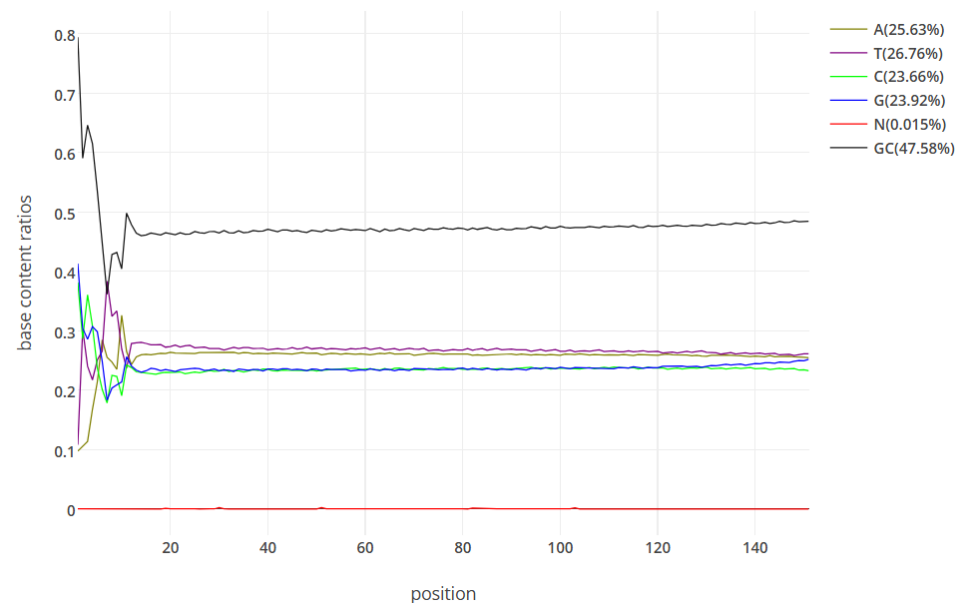
### Filtering result

|                         |                          |
|-------------------------|--------------------------|
| reads passed filters:   | 49.163226 M (97.776222%) |
| reads with low quality: | 1.080906 M (2.149715%)   |
| reads with too many N:  | 11.724000 K (0.023317%)  |
| reads too short:        | 25.516000 K (0.050746%)  |



与fastqc类似

去之前/去之后



# 测序质量相关的其他问题

- 取样：新鲜
- 多送：组织、细胞、RNA等
- 快送：快递、干冰、冰袋
- 数据产出：均匀
- 批次效应：人、时间、flowcell、lane
- 靠谱的合作伙伴